협업 필터링 추천 시스템을 위한 앙상블 기반 데이터 임퓨테이션 방법

An Ensemble-based Data Imputation Method for Collaborative Filtering Recommender Systems

저자 (Authors)
김형욱, 하지운, 김상욱
Hyung-ook Kim, Jiwoon Ha, Sang-Wook Kim

출처 (Source)
한국정보과학회 학술발표논문집 , 2015.06, 265-267(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06394765

APA Style
김형욱, 하지운, 김상욱 (2015). 협업 필터링 추천 시스템을 위한 앙상블 기반 데이터 임퓨테이션 방법. 한국정보과학회 학술발표논문집, 265-267

협업 필터링 추천 시스템을 위한 앙상블 기반 데이터 임퓨테이션 방법*

김형욱 ○, 하지운, 김상욱**

한양대학교 컴퓨터·소프트웨어학과

{kho0810, oneofus, wook}@agape.hanyang.ac.kr

An Ensemble-based Data Imputation Method for Collaborative Filtering Recommender Systems

Hyung-ook Kim○, Jiwoon Ha, Sang-Wook Kim**

Dept. of Computer & Software, Hanyang University

요 약

협업 필터링 기반 추천 방법은 데이터 희소성에 의해 추천의 정확도가 저하되는 문제가 있다. 이에 따라, 데이터 희소성을 해결하기 위한 데이터 임퓨테이션 방법들이 연구되어왔다. 그러나 기존 방법들은 채워지는 값의 정확도를 고려하지 않고 데이터 임퓨테이션을 수행하여, 정확도가 떨어지는 값들이 채워짐으로 인해 추천 정확도를 저하시키는 문제가 있었다. 따라서 본 논문에서는 앙상블에 기반하여 정확도가 높을 것으로 예측되는 값만을 채우는 새로운 데이터 임퓨테이션 방법을 제안한다. 실험 결과, 제안하는 방법이 기존 방법에 비해 향상된 추천 정확도를 보인다.

1. 서론

온라인 쇼핑몰에서는 아이템들에 대한 특정 유저의 선호도를 예측하고, 이 선호도를 바탕으로 해당 유저에게 일정 개수의 아이템을 추천하는 추천 시스템을 사용한다. 이러한 방식의 추천 시스템을 top-n 추천 시스템이라 하며, 선호도 예측의 대상이 되는 유저와 아이템을 각각 타겟 유저와 타겟 아이템이라 한다.

협업 필터링(collaborative filtering)은 추천 시스템 분야에서 널리 연구 및 사용되고 있는 방법으로, 적용 대상 데이터의 내용적 특성에 크게 구애 받지 않아, 다양한 데이터에 적용이 용이하다는 장점이 있다[1]. 협업 필터링 방법은 유사한 유저 또는 유사한 아이템의 정보를 이용하여 타겟 유저의 타겟 아이템에 대한 선호도를 예측한다.

협업 필터링 방법은 데이터 희소성(data sparsity)에 의해 추천 정확도가 저하되는 문제가 있다[2]. 데이터 희소성은 각각의 유저들이 모든 아이템 중 극히 일부분의 아이템에만 평점을 부여하여, 전체 평점의 수가 유저-아이템 쌍의 수에 비해 현저히 적은 상황을 의미한다. 유저들이 아이템들에 부여한 평점 정보는 행과 열이 각각 유저와 아이템으로, 각 셀(cell)의 값이 해당 유저가 해당 아이템에 부여한 평점을 의미하는 유저-아이템 행렬 R로 표현할 수 있다[1]. 일반적으로 여러 실제 데이터에서 유저-아이템 행렬의 밀도(density)는 10% 미만으로 알려져 있다. 예를 들어, 추천 시스템 연구에서 널리 사용되는 MovieLens 데이터와 Netflix 데이터의 경우 데이터의 밀도가 각각 6.30%와 1.17%에 불과하다. 이와 같이 데이터가 희소한 경우, 각 유저의 취향을 정확하게 파악하기 어렵기 때문에 유저간 또는 아이템간 유사도 계산의 정확도가 저하되고, 결과적으로 추천의 정확도 또한 저하된다[2].

데이터 희소성 문제를 해결하기 위해, 유저가 아직 평가하지 않은 아이템에 대해 선호도를 예측하여 셀을 채우는 데이터 임퓨테이션(data imputation) 방법이 제안되었다[2]. 데이터 임퓨테이션은 데이터 희소성 문제를 근본적으로 해결하고, 더 나아가 추천 정확도 향상에 도움을 주는 것으로 나타났다[2].

기존 데이터 임퓨테이션 방법들은 모든 빈 셀에 대해서 평점을 예측하고 그 값으로 셀을 채우는 방식으로 수행되었다[2]. 예측된 평점은 해당 아이템에 대한 유저의 실제 선호도와 정확히 일치하기는 어려우며, 이로 인해 각 셀에 예측된 평점과 실제 선호도 간에는 오차가 발생하게 된다. 기존 방법들은 이러한 오차를 고려하지 않고 모든 빈 셀을 채워, 예측된 평점과 실제 선호도 간의 오차가 큰 값도 채워지는 문제가 있다.

채워지는 평점이 실제 선호도와의 오차가 클 경우, 유저들의 취향을 잘못 파악하게 되어 유저간 또는 아이템간의 유사도 계산의 정확도가 저하된다. 이로 인해 타겟 유저와 실제로 유사하지 않은 유저들, 타겟 아이템과 실제로 유사하지 않은 아이템들을 통해 평점을 예측하게 되어 예측의 정확도 또한 낮아진다.

만일, 데이터 임퓨테이션을 통해 값을 채울 때, 정확도가

* 본 연구는 (1) 미래창조과학부 및 정보통신기술진흥센터의 대학ICT 연구센터육성 지원사업 (IITP-2015-H8501-15-1013), (2) 중소기업청에서 지원하는 산학연협력 기술개발사업 (No. C0191469), (3) 미래창조과학부의 재원으로 한국연구재단 (NRF-2014R1A2A1A10054151)의 지원을 받아 수행됨.

** 교신저자

[1] 유저가 아이템에 부여한 평점이 없는 경우 해당 셀은 비어있게 된다.

높은 값들만을 채울 수 있다면, 이러한 문제를 최소화하여 데이터 임퓨테이션의 효과를 더욱 극대화 시킬 수 있을 것이다. 그러나 데이터 임퓨테이션을 통해 채우고자 하는 셀에는 값이 존재하지 않으므로, 실제 해당 유저의 해당 아이템에 대한 선호도 대비 예측된 선호도의 정확도를 계산할 수 없다. 따라서, 실제 정확도를 계산할 수 없는 상황에서 별도의 접근 방법을 통해 정확하게 예측할 것으로 추정되는 셀을 찾아내는 방법이 필요하다. 이에 따라, 본 논문에서는 앙상블에 기초하여 상대적으로 정확하게 예측될 것으로 추정되는 셀을 찾고, 해당 셀들에 데이터 임퓨테이션을 수행하는 방법을 제안한다.

제안하는 방법의 우수성을 검증하기 위해, 본 논문에서는 MovieLens 데이터를 이용하여 top-n 추천을 수행하였다. 실험 결과, 제안하는 방법이 기존 방법보다 정확한 추천을 제공함을 보였다.

2. 제안하는 방법

본 논문에서는 채워지는 평점의 정확도를 고려하기 위해서 앙상블 방법을 사용한다. 앙상블은 데이터 마이닝에서 분류 문제(classification)와 예측 문제(prediction)의 정확도를 높이기 위해서 널리 사용되는 방법이다[3]. 일반적으로 앙상블은 여러 가지 방법으로 분류 혹은 예측을 수행하여 얻어진 결과들에 대해 다수결 또는 평균을 이용하여 최종 레이블(label) 혹은 값을 도출한다. 즉, 앙상블은 여러 방법에서 유사한 결과가 도출될수록 해당 결과가 정확할 것으로 간주한다고 볼 수 있다. 이 점에 착안하여 본 논문에서는 하나의 빈 셀에 대해 서로 다른 평점 예측 방법들을 각각 사용하여 평점을 예측하고, 해당 평점들 간에 차이가 작을수록 해당 평점이 정확할 가능성이 높을 것이라고 가정한다. 이를 바탕으로, 각각의 방법에 의해 예측된 평점 간의 차이가 작은 순으로 채울 셀을 선택하고, 해당 셀들을 앙상블을 이용하여 예측한 평점으로 채우는 앙상블 기반 데이터 임퓨테이션 방법을 제안한다.

제안하는 방법에서는 유저 기반 추천 방법[4]과 아이템 기반 추천 방법[5]의 앙상블을 통해 데이터 임퓨테이션을 수행한다. 두 추천 방법은 일반적으로 가장 널리 사용되는 협업 필터링 방법이다[1]. 유저 기반 추천 방법은 타겟 유저와 다른 유저들과의 유사도 계산을 통해 가장 유사한 취향을 가진 이웃 유저 k명을 찾고, k명 이웃 유저의 타겟 아이템에 대한 평점을 이용하여 타겟 유저의 타겟 아이템에 대한 평점을 예측한다. 아이템 기반 추천 방법은 타겟 아이템과 유사한 이웃 아이템을 k개 찾고, k개의 이웃 아이템에 대한 타겟 유저의 평점을 이용하여 타겟 유저의 타겟 아이템에 대한 평점을 예측한다.

제안하는 방법의 흐름은 다음과 같다. 첫 째, 유저 기반 추천 방법과 아이템 기반 추천 방법을 이용하여 모든 빈 셀에 대해서 평점을 예측한다. 둘 째, 비어있는 각 셀에 대해 수식 (1)의 예측 오차를 계산한다. 이 때, 유저-아이템 행렬 R의 빈 셀 R_{ui}에 대해서 두 추천 방법을 통해 예측된 평점을 각각 P_{ui}^{U}, P_{ui}^{I}라 한다. 여기서 u, i는 각각 해당 셀에 해당하는 유저와 아이템을 나타낸다.

$$\left|P_{ui}^{U} - P_{ui}^{I}\right| \quad (1)$$

셋 째, 모든 빈 셀에 대해 예측 오차가 계산되면, 예측 오차의 오름차순으로 셀들을 정렬한다. 넷 째, 정렬된 빈 셀들에서 상대적으로 정확하게 예측할 수 있는 상위 γ %의 빈 셀을 선택한다. 이는 예측된 값들 간의 차이가 작을수록 정확할 가능성이 높을 것이라는 앙상블의 가정에 기반 한 것이다. 이 때, γ는 임퓨테이션 비율로서 γ가 커질수록 유저-아이템 행렬의 밀도는 증가하지만 채워지는 평점들이 실제 선호도와 오차가 커지게 된다. 마지막으로, 채울 셀들을 선택한 이후에는 선택된 셀들에 대해서 P_{ui}^{U}와 P_{ui}^{I}의 평균값을 채운다.

3. 성능 평가

제안하는 방법의 우수성을 보이기 위해 MovieLens 데이터를 사용하였다. 실험에 사용한 MovieLens 데이터는 유저는 943 명, 영화는 1,682 개다. 모든 유저가 영화에 부여한 평점은 총 100,000 개이고, 각 유저는 최소 20 개 이상의 영화에 평점을 부여하였다. 평점은 1점과 5점 사이의 정수로 부여된다. 정확도 검증을 위하여 100,000 개의 평점을 4:1의 비율로 training set과 test set으로 나누어서 5번의 교차 검증(cross validation)을 수행하였다.

[표 1] Top-n 추천에서의 γ의 영향력.

추천 방법	임퓨테이션 비율	정밀도 (top 5)	정밀도 (top 10)	정밀도 (top 15)	정밀도 (top 20)	정밀도 (top 25)	재현율 (top 5)	재현율 (top 10)	재현율 (top 15)	재현율 (top 20)	재현율 (top 25)
유저 기반	0%	0.0012	0.0020	0.0032	0.0044	0.0060	0.0008	0.0022	0.0066	0.0134	0.0236
	20%	0.0786	0.0728	0.0674	0.0622	0.0576	0.0824	0.1392	0.1848	0.2202	0.2512
	40%	0.0588	0.0580	0.0544	0.0506	0.0468	0.0628	0.1170	0.1594	0.1936	0.2222
	60%	0.0548	0.0556	0.0524	0.0486	0.0450	0.0554	0.1090	0.1506	0.1832	0.2106
	80%	0.0600	0.0562	0.0526	0.0480	0.0450	0.0624	0.1112	0.1516	0.1808	0.2098
	100%	0.0752	0.0678	0.0604	0.0548	0.0502	0.0650	0.1106	0.1448	0.1752	0.1998
아이템 기반	0%	0.0410	0.0416	0.0396	0.0386	0.0374	0.0302	0.0602	0.0852	0.1140	0.1366
	20%	0.0774	0.0700	0.0656	0.0608	0.0570	0.0580	0.1060	0.1458	0.1790	0.2092
	40%	0.0612	0.0564	0.0536	0.0496	0.0460	0.0550	0.0992	0.1398	0.1738	0.2028
	60%	0.0542	0.0528	0.0500	0.0464	0.0432	0.0502	0.0926	0.1326	0.1642	0.1948
	80%	0.0526	0.0500	0.0488	0.0450	0.0420	0.0488	0.0898	0.1314	0.1628	0.1904
	100%	0.0632	0.0580	0.0544	0.0508	0.0466	0.0506	0.0898	0.1252	0.1554	0.1806

[표 2] 제안하는 방법과 기존 방법들의 top-n 추천 정확도 비교.

추천 방법	임퓨테이션 방법	정밀도 (top 5)	정밀도 (top 10)	정밀도 (top 15)	정밀도 (top 20)	정밀도 (top 25)	재현율 (top 5)	재현율 (top 10)	재현율 (top 15)	재현율 (top 20)	재현율 (top 25)
유저 기반	베이스라인	0.0012	0.0020	0.0032	0.0044	0.0060	0.0008	0.0022	0.0066	0.0134	0.0236
	유저 기반	0.0042	0.0086	0.0138	0.0190	0.0216	0.0024	0.0148	0.0338	0.0626	0.0876
	아이템 기반	0.0832	0.0718	0.0650	0.0604	0.0556	0.0678	0.1116	0.1510	0.1864	0.2148
	유저+아이템 기반	0.0752	0.0678	0.0604	0.0548	0.0502	0.0650	0.1106	0.1448	0.1752	0.1998
	제안하는 방법	**0.0786**	**0.0728**	**0.0674**	**0.0622**	**0.0576**	**0.0824**	**0.1392**	**0.1848**	**0.2202**	**0.2512**
아이템 기반	베이스라인	0.0410	0.0416	0.0396	0.0386	0.0374	0.0302	0.0602	0.0852	0.1140	0.1366
	유저 기반	0.0046	0.0102	0.0140	0.0164	0.0184	0.0028	0.0154	0.0310	0.0496	0.0690
	아이템 기반	0.0788	0.0722	0.0646	0.0580	0.0540	0.0638	0.1134	0.1468	0.1770	0.2068
	유저+아이템 기반	0.0632	0.0580	0.0544	0.0508	0.0466	0.0506	0.0898	0.1252	0.1554	0.1806
	제안하는 방법	0.0774	0.0700	0.0656	0.0608	0.0570	0.0580	0.1060	0.1458	0.1790	0.2092

데이터 임퓨테이션 결과가 추천 정확도에 미치는 영향을 검증하기 위해, 기존 방법과 제안하는 방법으로 각각 데이터 임퓨테이션 수행 후, 기존에 널리 사용되는 추천 방법인 유저 기반 추천 방법[4]과 아이템 기반 추천 방법[5]을 통해 top-n 추천을 수행하고, 그 정확도를 정밀도(precision)와 재현율(recall)을 통해 검증하였다.

먼저, 임퓨테이션 비율 γ의 영향력을 알아보기 위한 실험을 수행하였다. 표 1은 유저 기반 추천 방법과 아이템 기반 추천 방법으로 top-n 추천을 했을 때, γ를 변화시키며 정밀도와 재현율을 측정한 것이다. 그 결과, 유저 기반 추천 방법과 아이템 기반 추천 방법으로 top-n 추천을 했을 때 모두 γ = 20% 일 때가 정밀도와 재현율이 최대이고, γ가 그 이상 증가하면 정밀도와 재현율이 모두 감소하는 것을 알 수 있다. 이는 부정확한 평점들을 채우게 될 경우, 데이터 희소성은 완화되더라도, 결과적으로 추천의 정확도를 저하시키게 됨을 의미한다.

다음으로, 제안하는 방법의 우수성을 보이기 위해 기존 방법들과 마찬가지로, 채워지는 값의 정확도를 고려하지 않고 모든 빈 셀을 (1) 유저 기반 추천 방법을 이용하여 채우는 방법, (2) 아이템 기반 추천 방법을 이용하여 채우는 방법, 그리고 (3) 유저 기반 추천 방법과 아이템 기반 추천 방법의 평균으로 채우는 방법과 제안하는 방법 [2] 의 정확도를 비교하였다. 또한, 데이터 임퓨테이션을 수행하지 않고 top-n 추천을 수행한 결과를 베이스라인으로써 함께 비교하였다.

표 2는 기존 방법과 제안하는 방법에 대해서 각각 유저 기반 추천 방법과 아이템 기반 추천 방법으로 정밀도와 재현율을 측정한 것이다. 실험 결과, 제안하는 방법은 기존 방법들 중 가장 정확도가 높은 방법인 아이템 기반 추천 방법을 이용하여 모든 빈 셀에 데이터 임퓨테이션을 수행하고 유저 기반 추천 방법으로 top-n 추천을 수행한 방법에 비해 정밀도에서 최대 3.69%, 재현율에서 최대 24.73% 높은 정확도를 보이는 것으로 나타났다.

아이템 기반 추천 방법을 이용하는 데이터 임퓨테이션을 수행하고 유저 기반 추천 방식으로 top-n 추천을 하는 기존 방법의 경우, 제안하는 방법에 비해 top-5 추천에서 정밀도가 더 높게 나타난다. 그러나 표 2를 보면, 제안하는 방법을 통해 데이터 임퓨테이션을 수행한 경우, 추천하는 아이템의 개수가 증가할수록 기존 방법에 비해 정밀도가 높아진다. 또한, 재현율의 경우, 기존 방법들에 비해 확연히 높은 정확도를 보였다. 이러한 결과를 종합했을 때, 본 논문에서 제안하는 방법을 통해 기존 방법들보다 정확도 높은 top-n 추천이 가능함을 알 수 있다.

4. 결론

본 논문에서는 빈 셀에 채우려고 하는 값의 정확도를 고려하여 선택적으로 빈 셀을 채우는 데이터 임퓨테이션 방법을 제안하였다. 정확도를 예측하기 위해 데이터 마이닝 분야에서 널리 사용되는 앙상블 방법을 사용하였다. 실험을 통해 전체 빈 셀에 데이터를 채우는 것에 비해 정확도가 높을 것으로 예상되는 일부 셀만을 채우는 것이 추천 정확도를 향상시키는 것을 밝혔다. 또한, 기존 방법들에 비해 제안하는 방법이 높은 추천 정확도를 보이는 것을 보였다.

본 논문의 결과를 바탕으로, 더욱 다양한 추천 방법을 앙상블 과정에서 이용하여 앙상블 기반 데이터 임퓨테이션 방법을 더욱 발전 시킬 수 있을 것으로 기대한다.

참고 문헌

[1] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 17, No. 6, pp. 734-749, 2005.

[2] Y. Ren et al., "The Efficient Imputation Method for Neighborhood-based Collaborative Filtering," *in Proc. of the 21st ACM International Conference on Information and Knowledge Management*, pp. 684-693, 2012.

[3] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufman, 2012.

[4] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *in Proc. of the 14th Conference on Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.

[5] B. Sarwar et al., "Item-based Collaborative Filtering Recommendation Algorithms," *in Proc. of the 10th ACM International Conference on World Wide Web*, pp. 285-295, 2001.

[2] 제안하는 방법에서 γ은 20%로 설정하였다.